On the use of Large Language Models to Detect Brazilian Politics Fake News

Marcos P. S. Gôlo[1] , Adriel L. V. Mori[2] , William G. Oliveira[2] , Jacson R. Barbosa[2] , Valdemar V. Graciano-Neto[2] , Eliomar A. de Lima[2] , Ricardo M. Marcacini[1]

[1]Institute of Mathematics and Computer Sciences - University of São Paulo (USP)

[2]Federal University of Goiás (UFG)

{marcosgolo,ricardo.marcacini}@usp.br

{adrielmori,william.william}@discente.ufg.br

{jacson_rodrigues,valdemarneto,eliomar}@ufg.br

Abstract. *Machine learning methods are proposed to mitigate the spread of fake Brazilian news about politics so as not to harm society. Supervised algorithms are explored, requiring labeled news to train. However, labeling a high volume of news can be complex, onerous, time-consuming, error-prone, and costly. Hence, large language models (LLMs) have been used to detect fake news once LLMs are unsupervised methods that can play the role of classifiers. Most fake news detection studies explore the OpenAI LLMs (require payment) and lack an empirical evaluation with other LLMs. However, several open-source models obtain comparative and state-of-the-art (SOTA) results. We highlight that these models have yet to be explored in detecting fake Brazilian news about politics, which is crucial as it directly impacts society. In this sense, we propose a new dataset for detecting fake Brazilian news about politics and an empirical evaluation of open-source LLMs and OpenAI LLMs. In our results, the LLM from Google (Gemma) outperformed the other six LLMs, including GPT-4, proving to be the most promising model for detecting fake news about Brazilian politics.*

1. Introduction

Fake news are news written to deceive and manipulate users [Souza 2023]. The rapid spread of fake news has been facilitated by the popularization of the Internet and social networks, which can cause harm to society [Rohera et al. 2022]. For instance, the spread of fake news about COVID-19 vaccines influenced people to refuse the vaccine, which harmed the health sector and even other people. In Brazil, politics is one of the main topics impacted by fake news. Data presented in [Santos 2022] show that social networks are widely used to discuss and present political content in Brazil, with potential to rapidly spread real or misleading information [Santos 2022]. Given the volume of news, machine learning methods are proposed to detect fake news and mitigate its impact on society. One of the most common forms to solve this problem is through binary supervised learning algorithms [Santos 2022, Souza 2023].

In binary supervised learning, algorithms require extensively labeled datasets for the algorithm training stage [Mishra et al. 2022]. One-class learning methods have been proposed to mitigate this limitation but still require fake news labeling [Gôlo et al. 2023].

However, labeling a large volume of news can be complex, time-consuming, and costly [Souza 2023]. In this sense, unsupervised methods based on machine learning have been used to detect fake news, especially methods based on large language models (LLMs) [Benny 2023, Chang et al. 2023]. LLMs are pre-trained models on a corpus with trillions of words capable of generating sentences from an input text [Hu et al. 2024]. LLMs can be considered unsupervised when used as pre-trained models to generate text and solve tasks such as fake news detection. Thus, LLMs have been used in the literature to detect fake news by receiving the news excerpts as input and generating an output representing the classification for that news, such as true or false [Teo et al. 2024].

Recent studies have explored LLMs to detect fake news using different strategies just with OpenAI LLMs. Thus, the studies explored only a single type of LLM [Qu et al. 2024, Hu et al. 2024, Pelrine et al. 2023]. Although the LLM for fake news detection studies have achieved impressive results in terms of accuracy, they do not explore the Brazilian political context, where extreme polarization and the rapid dissemination of disinformation, particularly during electoral periods, present unique challenges [Junkert 2022]. Modeling and detecting fake news in this context is crucial, as it can directly impact public opinion and political stability. The scarcity of studies focused on Brazil using dedicated political datasets leaves a significant gap in the literature.

This paper proposes an empirical evaluation of different LLMs to detect Brazilian fake news about politics. To achieve this research goal, we collected and constructed a dataset of Brazilian news about politics. In this sense, we collected real and fake news to empirically evaluate the comparison between the LLMs. We used seven LLMs as classifiers (fake news detectors), explaining why a news item is fake or real, and generating news embeddings to analyze the generated representations. Based on the experiments conducted, we answered the following research questions (RQ):

1. **RQ 1:** Which large language model obtains the best classification performance to detect Brazilian Fake News about Politics?
2. **RQ 2:** Can large language models explain why the Brazilian news of politics is fake or real, presenting misinformation or veracity features of the news?
3. **RQ 3:** Which large language model generates the best representation to better separate real from fake news and serve as an unsupervised representation method?

2. Related Work

Recent studies have explored LLMs for fake news detection. In particular, [Hu et al. 2024] utilized two real-world datasets to evaluate the performance of their models: Weibo21 [Nan et al. 2021] and GossipCop [Shu et al. 2020]. The prompt strategy involved crafting specific questions for the LLM to gain insights into news analysis, effectively leveraging the model's comprehension capabilities. The study employed a few-shot approach, where the model is trained on a few examples, demonstrating the potential for efficient learning with limited data. Only GPT-3.5 was employed, and the model obtained 0.784 and 0.790 of f1-score, outperforming baseline methods (Small Language Models), highlighting the value of integrating LLMs in enhancing smaller models.

Another significant contribution is [Pelrine et al. 2023], which adopted GPT-4 as the main tool. The authors examined three datasets to evaluate the effectiveness of their approach: the widely-used LIAR [Wang 2017], CT-FAN-22 [Köhler et al. 2022]. A new dataset, LIAR-New, was used in a study where the prompt strategy was carefully designed to maximize the model's ability to identify misinformation patterns, implementing both zero-shot and few-shot approaches. These approaches tested the model with little prior exposure to the specific data type. For performance comparison, the study employed multiple language models, but when focusing on LLMs, only GPT-3 and GPT-4 were used, which are variations of OpenAI. GPT-4 obtained an F1-score of 87%, significantly outperforming previous models.

[Qu et al. 2024] proposed a quantum multimodal fusion model that combined textual and visual features using a quantum convolutional neural network (QCNN). This model was tested on two datasets, achieving an accuracy of 87.9% and 84.6%, highlighting its robustness against quantum noise and the efficiency of quantum encoding in improving the accuracy of fake news detection. The study utilized the text-davinci-003 from OpenAI, further emphasizing the effectiveness of advanced LLMs in this domain.

[Li et al. 2024] introduced FactAgent, an agentic approach that utilizes an LLM for fake news detection. FactAgent distinguishes itself by employing a structured workflow that simulates the behavior of human experts in verifying news claims. This model was evaluated on three well-known datasets: Snopes, PolitiFact, and GossipCop. FactAgent operates in a zero-shot setting, meaning it does not require prior training on labeled datasets. Instead, it leverages the internal knowledge of LLMs and external tools, such as search engines, to assess the veracity of news articles. FactAgent achieved impressive results, with accuracy scores of 88%, 83%, and 75% on PolitiFact, GossipCop, and Snopes, respectively. The authors use the GPT-3.5 LLM. These results underscore the model's capability to integrate LLM reasoning with external evidence, providing a highly adaptable and efficient solution for fake news detection without annotated training data.

The studies presented state-of-the-art results for fake news detection. On the other hand, we highlight two gaps. The first gap is that they do not explore the Brazilian political context, in which the rapid spread of fake news presents challenges [Junkert 2022]. The second is using a single type of LLM in a scenario that can be solved with different open-source LLMs. In addition, we highlight the need for more empirical studies on which is the best LLM to detect fake news, which would benefit the literature when choosing the LLM to solve the fake news detection problem. Most studies explore GPT as LLM, and [Qu et al. 2024] uses davinci-003. All models are from Openai, i.e., they are models that can only be used upon payment. Thus, in the next section, we present an empirical evaluation of different LLMs for detecting fake news in Brazilian politics.

3. Research Method

This section presents the research method adopted in this study. Our goal is to present and make available our dataset for political fake news detection and demonstrate which large language model performs better in detecting these fake news types. The next sections present the dataset used in the experimental evaluation, our strategy to use LLMs as classifiers, and experimental settings. The codes and dataset are publicly available.[1].

[1]https://github.com/GoloMarcos/LLM4BrazilianFakeNews.git

3.1. Dataset

To create dataset, we chose three hundred relevant news from Brazil's recent presidential elections (2018 and 2022) and other relevant news about politicians. Concerning the real news, they were ready for analysis. However, fake news requires some attention since the majority of fake news is news that explains why certain content is fake. We are not interested in this type of news. Therefore, we collect the content, in its entirety, from news that explains why certain content is fake. The data highlights mainstream articles and social media posts from the past five years involving governmental news and political figures focusing, for instance, on Presidents and the minister of the federal supreme court.

We emphasize that there is another dataset of news about Brazilian politics that detects fake news presented in the studies [Souza 2023] and [Gôlo et al. 2023]. On the other hand, the dataset was built with a crawler and without collecting the fake news in its entirety, i.e., the dataset has news that states about the fake news instead of the fake news itself. This dataset's characteristics bias the model in the learning to detect news of this type, which harms the model if the model is used in production in the real world.

The news sources of the dataset are G1, Correio Brasiliense, UOL, and Aos Fatos. They are exceptionally respected news sources in the Brazilian context. For preventive disinformation measures, these news outlets have tools to guide journalists in deciding whether the news is fake, which helps the public avoid sharing misinformation on the Internet. The news distribution is 152 real news and 148 fake news, and for the fake news, there are 98 videos, 41 images, six audios, and three texts. Figure 1 presents information about our gold standard set. In the news, with video, image, or audio sources, we collect the text involved in the news, for instance, the title or description of the post. We also present the word cloud for our dataset (Figure 2).

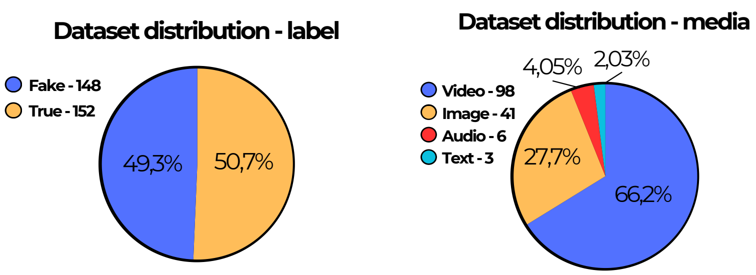

Figure 1. Dataset distribution based on true or fake labels and the media type.

About the word cloud, some words are used in several news, such as those that show discontent, criticality, politicians' names, mainly the leaders of their respective political parties, and the names of the government justice bodies, especially the most important one, the federal supreme court. Figure 2 shows the most common words in the dataset text news. Most parts are words used in a political context, and that reflects the main proposal of our project and our commitment to choosing unique news that is in a political context, the ones that show the fake content adequately, and the body availability of the news article itself so that it can be used in the dataset this requirements provides us a high-quality dataset to training the LLMs.

Figure 2. The word cloud of our collected dataset.

3.2. LLMs to Detect Brazilian Politics Fake News

Large Language Models (LLMs) are trained by predicting the next word of a sentence given a sequence of previous words. Given a sequence of words in a textual document $d = (w_1, w_2, ..., w_v)$, the LLM first generates a corrupted version $\hat{d}$ of the sequence, where the last word (w_v) is selected to be replaced by a special token called [MASK]. Thus, the objective function is to reconstruct the masked token $\bar{d}$ from $\hat{d}$, i.e., to predict the next word of a sentence [Liu et al. 2023]. The model is trained by maximizing the probability of a given sequence of tokens conditioned on the context. Define w_v as the target variable y. Therefore, the probability can be defined as:

$$P(y|d) = P(y|w_1, w_2, ..., w_{v-1}), \tag{1}$$

where $w_1, w_2, ..., w_{v-1}$ are the tokens in the context sequence and v is the current position. Then, the conditional probability can be decomposed into [Chang et al. 2023]:

$$P(y|d) = \prod_{v=1}^{V} P(y_v|w_1, w_2, ..., w_{v-1}), \tag{2}$$

where V is the size of the sequence of tokens.

After the training of LLMs, we can use it to generate text from input texts since it is able to predict the next word. This allows our study to send news to the LLM and ask it to return if the news is fake or real. It is worth noting that the LLMS are trained with billions of tokens, which makes them a powerful tool for generating text from an input [Chang et al. 2023]. Different LLMs have been proposed in the last 3 years, and they have differences that generate advantages and disadvantages for each model [Zhao et al. 2023].

The main evaluation criterion is the family of the LLM. We explore in our experiment five families of LLMs: OpenAI[2] (GPT) [OpenAI 2024], Meta (LLaMa)

[2] https://openai.com/

[Meta 2024], Microsoft (Phi) [Microsoft 2024], Google (Gemma) [Gemma 2024], and Alibaba Cloud (Qwen) [Alibaba 2023]. In the second one the focus is on the version of the LLM and its number of parameters. We explore the GPT 4 and 3.5-turbo, Llama 3 (8 billion of parameters), Phi 3 (14 billion of parameters), Gemma 2 (27 billion of parameters), Qwen 1(32 billion of parameters), Qwen 2 (7 billion of parameters). We choose models that use less than our 24GB of GPU (see experimental setting for details).

We highlight other differences between the LLaMA and GPT models. First, regarding the Multilingualism, GPT tends to perform poorly in languages other than English, but LLaMA is designed to excel in multiple languages, which is better for our Brazilian scenario. Second, GPT offers models with a significantly larger token limit than LLaMa. Third, LLaMa is often considered faster and more resource-efficient than GPT since GPT is larger than LLaMa considering the number of parameters. Fourth, the usability. LLaMA is open-source and more accessible. However, GPT needs to be paid [Roumeliotis et al. 2024]. We show our pipeline in Figure 3.

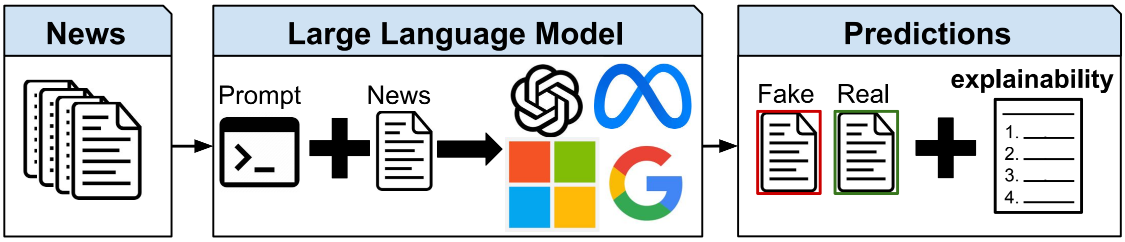

Figure 3. Our pipeline proposal. First, we collect the news; second, we use the different large language models to predict whether the news is fake or real with explainability.

3.3. Experimental Settings

We use the following prompt to evaluate the LLMs: "*You are a fack-checker. Answer whether the following news is fake or real. Your answer should be only the word fake or real. Follow the news:* ***news***. *Remember, your answer should be only the word fake or real*", in which **news** is the text content of the news. We use the *Ollama* library to run the open-source models[3]. We pay for the GPT model and use the *OpenAI* library. We execute the experimental evaluation in a machine with an Ubuntu 24 computer with an i9-14900KF CPU, RTX A5000 (24 GB RAM), and 128 GB RAM.

Since the LLMs work in an unsupervised way and do not require a training set, we do not adopt any cross-fold validation process. All the news (real and fake) are used as a test set. In addition, we use precision, recall, and f_1-score for fake and real classes, and the f_1-macro as the evaluation measures.

4. Results and Discussion

Table 1 presents the results of our study. We present the precision, recall, and f_1 for real and fake classes. We also present the f_1 macro for the seven models explored. Higher values are in bold (best models). Gemma 2 was the best model since it obtained the higher f_1 macro. GPT 3.5 turbo obtained the second-best results, followed by Qwen 1 and Qwen

[3] `https://ollama.com/`

2. Phi 3 obtains the worst f_1 macro, followed by Llama3 and GPT 4. We highlight open-source models, such as Gemma, that, compared to GPT4, which has state-of-the-art results in other tasks, outperformed the paid model.

Table 1. Large Language Models results for Brazilian politics fake news detection. Each line represents one model, and each column represents one metric. In brackets next to the models' names, we add the number of parameters of each open-source model. The best results are in bold.

Models	fake			real			f_1-macro
	p	r	f_1	p	r	f_1	
Llama 3 (8b)	0.93	0.26	0.41	0.58	0.98	0.73	0.57
Gemma 2 (27b)	0.95	0.83	**0.89**	0.85	0.96	**0.90**	**0.90**
Phi 3 (14b)	**0.97**	0.23	0.37	0.57	**0.99**	0.72	0.55
Qwen 1 (32b)	0.80	0.82	0.81	0.82	0.80	0.81	0.81
Qwen 2 (7b)	0.89	0.60	0.72	0.70	0.93	0.80	0.76
Gpt 3.5 turbo	0.80	0.89	0.85	0.88	0.79	0.83	0.84
Gpt 4	0.61	**0.99**	0.75	**0.98**	0.38	0.54	0.65

Considering other metrics, Gemma 2 obtains the best f_1-score for fake and real classes, reflecting the higher f_1 macro. Different methods obtained better results for recall and precision. Phi 3 obtained a higher precision for the fake class and higher recall for the real class. GPT 4 obtains the best recall for the fake class and the best precision for the real class. However, these methods obtain the worst f_1 macro. Why?

These methods were good at identifying only one class, obtaining good results for only one metric in each class. GPT 4 was suitable for identifying the fake class since the method obtained a higher recall for the fake class and higher precision for the real class, i.e., the method does not predict fake news as real. However, GPT 4 obtains the worst precision for fake class and recall for real class, i.e., the method predicts real news as fake. Phi 3 had the same behavior but in the real class.

Next to the models' names, we add the parameters for each method. In this sense, we note that the open-source methods with more parameters, Gemma 2 with 25 billion and Qwen 1 with 32 billion, obtain better results than the methods with fewer parameters (Phi 3 with 14 billion, Llama 3 with 8, and Qwen 2 with 7). Even though methods with more parameters obtain the best results, it is worth highlighting that the number of parameters is not directly proportional to the result. This can be observed because the Phi 3 model has more parameters than the Llama 3 and Qwen 2 but obtained worse results.

In addition to using LLMs as text classifiers, we also explore them as explainability models since it generate text. The models can generate explanations for the news predictions, i.e., why the news is real or fake. We choose the follow fake news to analyze: ***Alexandre de Moraes e outros funcionários do governo ameaçaram processar criminalmente o advogado do Twitter no Brasil se ele não entregasse informações privadas e pessoais, incluindo números de telefone das pessoas e suas mensagens diretas pessoais!***. We choose the top 3 models to generate an textual explanability for this news:

- **Gemma 2**: *Here are three misinformation features:*

1. ***Vague Language:*** *The statement uses sweeping terms like "ameaçaram processar criminalmente" without specifying the exact nature of the threat or legal grounds for such action. This vagueness makes it difficult to assess the veracity of the claim.*
2. ***Lack of Context:*** *The statement doesn't provide crucial context, such as: (i) What information was specifically requested and why? (ii) Was there a legal basis for the request (e.g., a court order)? (iii) What was the response from the Twitter lawyer? Without this context, it's impossible to determine if the actions of Alexandre de Moraes and other government officials were appropriate or excessive.*
3. ***Potential for Sensationalism:*** *The statement focuses on the dramatic element of a potential criminal lawsuit against the Twitter lawyer. This could be an attempt to sensationalize the story and generate outrage without providing a balanced and accurate account.*

- **Gpt 3.5 turbo**: *Misinformation features (if judged as fake):*
  1. ***Outrageous and dramatic language:*** *the use of terms like "ameaçaram processar criminalmente" (threatened to criminally sue) can be a sign of sensationalism, which is a common tactic used in fake news.*
  2. ***Lack of credible sources:*** *the news does not provide any specific sources or evidence to support the claim that Alexandre de Moraes and other government officials made these threats.*
  3. ***Implausible scenario:*** *the idea that government officials would threaten legal action against a social media platform's lawyer in order to obtain personal information of users seems unlikely and raises questions about the credibility of the news.*
- **Qwen 1**: *Misinformation features:*
  1. ***Lack of credible sources:*** *The news does not cite any reputable sources or links to the original statements made by Alexandre de Moraes or government officials. Reliable reporting typically includes quotes or references to official documents, press releases, or interviews with those involved.*
  2. ***Sensational language:*** *The use of strong and emotive words like "ameaçaram" (threatened) and "criminalmente" (criminal charges) suggests a biased perspective and might be an attempt to incite fear or outrage without providing objective facts. Credible news reporting tends to maintain a neutral tone and presents information in a balanced manner.*
  3. ***Absence of context:*** *A true story would typically provide context, such as the reason behind any legal action being considered or the relevance of the requested information to an ongoing investigation. The absence of this context raises concerns about the accuracy and completeness of the claim.*

All models correctly identified some key misinformation features of the fake news. Some common features were identified by all three models, such as sensationalism, lack of credible sources, and lack of context. We highlight that this is the line of reasoning of humans. The superior electoral court published a news item that pointed out eye-catching titles and unknown websites or channels as misinformation features[4]. The Qwen

[4] https://www.tre-pr.jus.br/comunicacao/noticias/2023/Setembro/como-identificar-fake-news-na-duvida-nao-compartilhe-1

model cited these three features, which makes its explainability more generic and less informative. On the other hand, the Gemma model cited vague language as a misinformation feature, which makes sense since the news does not mention what type of criminal prosecution was applied. In addition, the GPT 3.5 turbo model also cited an implausible scenario as a key misinformation feature, which also makes sense since a search and seizure warrant is usually issued to obtain people's phones rather than a lawsuit.

It is also important that the model shows veracity features for real news. Therefore, we also present the explainability of the three models for the following real news: ***Com o voto do ministro Flávio Dino, a Primeira Turma do Supremo Tribunal Federal formou maioria nesta terça-feira (27) para manter decisão que derrubou uma condenação bilionária da Petrobras pela Justiça Trabalhista. O voto do relator, ministro Alexandre de Moraes, pela rejeição dos recursos apresentados por sindicatos foi seguido pela ministra Cármen Lúcia e por Dino. Como a Primeira Turma é composta por cinco ministros, os três votos já garantem maioria. Moraes afirmou que os sindicatos queriam rediscutir a tese fixada pelo Supremo, o que não é possível com o tipo de recurso apresentado. O julgamento ocorre no plenário virtual do Supremo, quando os ministros inserem seus votos no sistema eletrônico. O caso envolve a maior condenação trabalhista imposta à Petrobras, que poderia ter um impacto de ao menos R$ 17 bilhões para a estatal.***

- **Gemma 2**: *This news appears to be true. Here are three veracity features:*
  1. ***Attribution and Source:*** *The article clearly states its source as TV Globo, a reputable Brazilian news organization. It also names the author (Márcio Falcão) and provides a date and time stamp for publication.*
  2. ***Specific Details:*** *The article includes specific details about the case, such as the names of the ministers involved (Flávio Dino, Alexandre de Moraes, Cármen Lúcia), the amount of the potential fine (R$17 billion), and the nature of the legal dispute (a condemnation by the Labor Court).*
  3. ***Contextual Information:*** *The article provides context for the decision, explaining that it upholds a previous ruling by Minister Moraes suspending the fine. It also mentions the potential impact on Petrobras' finances.*
- **Gpt 3.5 turbo**: *Veracity features:*
  1. *The news source is TV Globo, a well-known and reliable media outlet.*
  2. *The text mentions specific details, such as the vote of Minister Flávio Dino and the formation of a majority in the First Chamber of the STF.*
  3. *The content of the news is consistent with the coverage of other media outlets on the same topic.*
- **Qwen 1**: *Veracity features:*
  1. ***Source attribution:*** *The news is attributed to Márcio Falcão from TV Globo, which indicates that it was written by a professional journalist and likely based on reliable information.*
  2. ***Specific details:*** *The article provides specific details such as the date of the decision (27/02/2024), names of the ministers involved (Flávio Dino, Alexandre de Moraes, and Cármen Lúcia), and the amount of the potential fine for Petrobras (R$ 17 billion).*
  3. ***Official context:*** *The reference to the "Supremo Tribunal Federal" (STF) and the description of the legal process involving the Primeira Turma suggests that the information is based on actual court proceedings, which are publicly accessible.*

All models correctly identified some key veracity features. Some of the most common features were identified by all three models, such as source and specific details. Once again, we highlight that these LLMs followed the line of human reasoning to verify whether a news item is real. The other feature identified by each model was different, which differentiates and ranks the explanations. The GPT model indicated consistency with other sources as a characteristic of real news. Consistency helps verify whether a news story is real, but it must be consistent with renowned sources rather than quantity. The GPT model did not mention this information in its explicability, which removes credibility from its third feature. On the other hand, the Gemma and Qwen models presented features with more credibility. Both cited the context of the news for public information and explanatory information that the fake news lacks and that the real news has.

Another critical point is the complementarity of the models. For instance, GPT and Qwen pointed out the lack of sources as a feature of misinformation and the source of the news as a feature of veracity. This fact also happened with Gemma due to the news's lack of context/information. Therefore, the models could generate explanations that made sense and were not hallucinatory.

In addition to comparing the LLMs classification performances, we performed another experiment to analyze the representations generated by the LLMs embeddings for the news. This analysis is interesting when unsupervised methods are used as text representation methods as input for other classifiers. Figure 4 presents two-dimensional projections of the embeddings considering each LLM in the Brazilian politics fake news dataset. We generated the representations using the t-Distributed Stochastic Neighbor Embedding (t-SNE) for the analysis [Van der Maaten and Hinton 2008]. All open-source models satisfactorily separate real and fake news. On the other hand, the OpenAI model had the worst performance, separating fewer news points. Among the models with the best results, we highlight the models that clustered the news more, such as Gemma. These models also obtained the best f_1-macro, possibly due to how they represent the news.

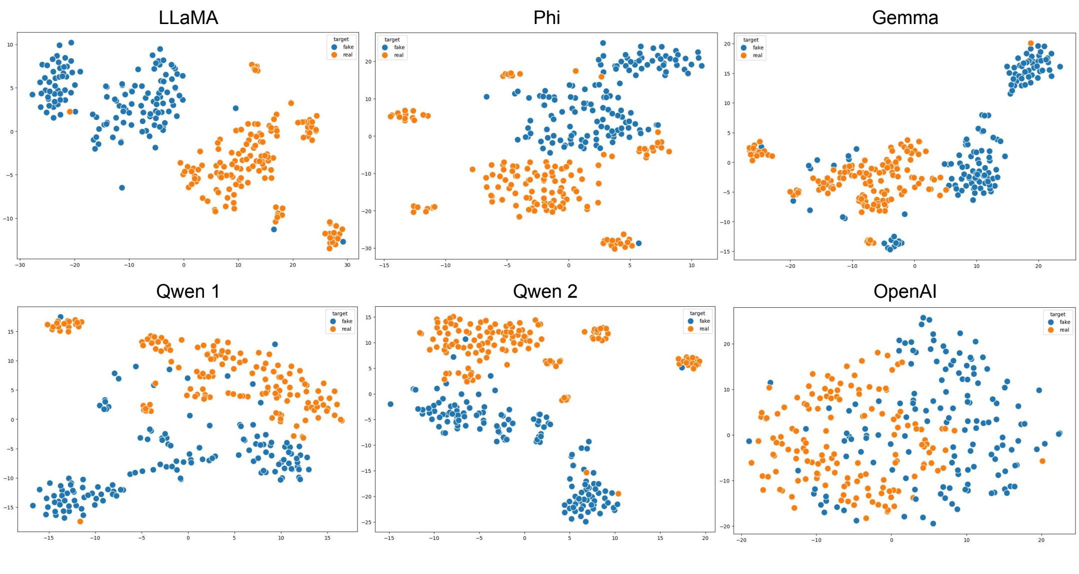

Figure 4. t-SNE (2D) of each LLM model. The colors indicate class real news (orange) and fake news (blue). Models that show less overlap between classes are more promising for fake news detection and representation.

5. Conclusions and Future Work

In this paper, we present a new dataset for detecting fake news about Brazilian politics and an empirical evaluation of LLMs on this dataset. LLMs represent the state-of-the-art for text mining tasks as well as for fake news detection. We explore the main LLMs from main international companies such as Gemma, Phi, Qwen, GPT, and LaMMa. We perform this empirical evaluation to answer the three research questions in the introduction.

Regarding **RQ 1 (best LLM performance)**, the Gemma model obtained the best LLM performance for detecting fake news about Brazilian politics. We highlight that Gemma is an open-source model, which is another advantage and incentive for its use. Regarding **RQ 2 (LLM explainability)**, the LLMs that obtained the best performances were able to generate explanations for why the news is fake or real. The models presented features of misinformation and veracity for the news related to the way a human verifies news. Regarding **RQ 3 (embedding)**, the Gemma and Qwen models generated the best representations since, in the two-dimensional plot, the models could separate fake and real news with less overlap and obtain a greater density for each news cluster.

Future directions indicate the collection of a larger dataset and a larger empirical evaluation. In addition to exploring different LLMs, different prompting strategies should be explored to enrich the empirical evaluation of LLMs. Lastly, two-step models (representation and classification) should be explored to evaluate the LLM's representations.

Acknowledgments

This work was supported by FAPESP (grant numbers 2023/10100-4 and 2019/07665-4), CAPES (grant number 88887.671481/2022-00), and LatAm Google Ph.D. Fellowship.

References

Alibaba (2023). Qwen technical report.

Benny, J. J. (2023). *Knowledge Informed Fake News Detection Using Large Language Models*. PhD thesis, University of Windsor (Canada).

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Gemma (2024). Gemma 2: Improving open language models at a practical size.

Gôlo, M. P. S., de Souza, M. C., Rossi, R. G., Rezende, S. O., Nogueira, B. M., and Marcacini, R. M. (2023). One-class learning for fake news detection through multimodal variational autoencoders. *Engineering Applications of Artificial Intelligence*.

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., and Qi, P. (2024). Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38.

Junkert, F. G. (2022). Fake news and the 2018 brazilian presidential election. In *The Rule of Law in Cyberspace*, pages 167–185. Springer.

Köhler, J., Shahi, G. K., Struß, J. M., Wiegand, M., Siegel, M., and Mandl, T. (2022). Overview of the CLEF-2022 CheckThat! lab task 3 on fake news detection. In *Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum*.

Li, X., Zhang, Y., and Malthouse, E. C. (2024). Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. (2023). Gpt understands, too. *AI Open*.

Meta (2024). The llama 3 herd of models.

Microsoft (2024). Phi-3 technical report: A highly capable language model locally on your phone.

Mishra, S., Shukla, P., and Agarwal, R. (2022). Analyzing machine learning enabled fake news detection techniques for diversified datasets. *Wireless Communications and Mobile Computing*, 2022(1):1575365.

Nan, Q., Cao, J., Zhu, Y., Wang, Y., and Li, J. (2021). Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

OpenAI (2024). Gpt-4 technical report.

Pelrine, K., Reksoprodjo, M., Gupta, C., Christoph, J., and Rabbany, R. (2023). Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv*.

Qu, Z., Meng, Y., Muhammad, G., and Tiwari, P. (2024). Qmfnd: A quantum multimodal fusion-based fake news detection model for social media. *Information Fusion*.

Rohera, D., Shethna, H., Patel, K., Thakker, U., Tanwar, S., Gupta, R., Hong, W.-C., and Sharma, R. (2022). A taxonomy of fake news classification techniques: Survey and implementation aspects. *IEEE Access*, 10:30367–30394.

Roumeliotis, K. I., Tselikas, N. D., and Nasiopoulos, D. K. (2024). Llms in e-commerce: a comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056.

Santos, R. L. d. S. (2022). *Detecção automática de notícias falsas em português*. PhD thesis, Universidade de São Paulo.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2020). Fakenewsnet: A data repository with news content, social context and spatiotemporal information for studying fake news on social media. *Big data*, 8:171–188.

Souza, M. C. d. (2023). *Detecção de notícias falsas usando poucos dados positivos rotulados*. PhD thesis, Universidade de São Paulo.

Teo, T. W., Chua, H. N., Jasser, M. B., and Wong, R. T. (2024). Integrating large language models and machine learning for fake news detection. In *2024 20th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, pages 102–107. IEEE.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11):2579–2605.

Wang, W. Y. (2017). ”liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Annual Meeting of the ACL*, pages 422–426.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.